Bioinformática ¿una ciencia sin científicos?

Biological computing science, a science without scientifists?

Roderic Guigó Serra

El Proyecto Genoma Humano ha catalizado una presencia sin precedentes de la investigación en biología en los medios de comuni Este impacto mediático no es gratuito. El conocimiento de la secuencia de nucleótidos del genoma humano y de la secuencia de aminoácidos de las proteínas codificadas en ese genoma tendrá, se dice, un impacto extraordinario en la medicina, la agricultura y muchos procesos industriales. Tendrá, en consecuencia, repercursiones económicas, sociales y quizás, incluso, políticas. En definit afectará profundamente nuestras vidas y es lógico que despierte nuestro interés.

The Human Genome Project has promoted an unprecedented presence of information on biological research in the media. This is not a gratuitous impact. It is widely believed that the accrued knowledge on human genome nucleotide sequences and on amino acid sequences of proteins codified by our genome will have an exceptional impact on medical sciences, agricultural sciences and many industrial processes. That is, it will cause financial, social and perhaps even political repercussions. In other words, it will deeply affect our lives, and thus is worthy of our interest.

Al margen de esa novedad mediática, la biología está sufriendo también una radical transformación en el mc que su práctica, como disciplina científica, se lleva a cabo. La biología, una ciencia tradicionalmente de la observación y la descripción, se está convirtiendo en una ciencia caracterizada por la generación de cantidad ingentes de información. La invención y el desarrollo de tecnologías diversas –alrededor de las cuales se arti esta disciplina científica que hemos dado en llamar *genómica*– son los responsables. No se trata sólo de la automatización y robotización de los métodos de secuenciación del DNA, que permiten la obtención de la secuencia completa de los genomas de los organismos vivos o la secuencia parcial de aquellos de entre los g codificados en este genoma, que se expresan en una determinada estirpe celular, y cuyos resultados son estos tan evidentes, sino también de la invención de las matrices de DNA (*DNA arrays*). Estas matrices permiten, particular, monitorizar la expresión simultánea de miles de genes en condiciones distintas; el aumento de la precisión de las técnicas de geles bidimensionales y de espectroscopia de masas, que permiten la caracteriza global de las proteínas en que estos genes son traducidos; las técnicas, como la de «dos híbridos» en levadur permiten inferir globalmente las interacciones entre estas proteínas, la complejidad de estas interacciones la sustenta los procesos de la vida, y la automatización y robotización de las técnicas de rayos X y de resonanci magnética nuclear, que están acelerando sustancialmente el descubrimiento de las estructuras tridimensional un gran número de estas proteínas.

Con la genómica, se ha producido un cambio de énfasis en el objeto de estudio de la biología molecular y ce del estudio de los genes y de las proteínas individuales pasamos al estudio de la acción sinérgica de decenas miles de esos genes y proteínas en la célula viva. El resultado inmediato de la aplicación de las técnicas gen es la obtención, casi automática, de cantidades inmensas de datos, de una magnitud insólita en la historia de biología. En este sentido, con la genómica, la biología se ha convertido en una ciencia de la información, tan la obtención de los datos genómicos primarios, como en su almacenamiento, análisis e integración; la inform desempeña, en consecuencia, un papel crucial; y una nueva disciplina científica, la bioinformática, en la intersección entre biología y computación ha emergido recientemente para hacer frente a las especificidades tratamiento de estos datos comporta.

La magnitud de la información que genera la investigación genómica es tal que, probablemente, supera la m de la información que genera la investigación en otras disciplinas científicas. No en vano, la vida es la forma compleja de organización de la materia que conocemos. En estos momentos, los ordenadores no clasificados uso civil más potentes del mundo (en Celera y en Oak Ridge National Laboratory, por ejemplo, con una capa de cálculo cercana a los 2 Teraflops, billones de operaciones por segundo) se encuentran ya dedicados a la investigación biológica, concretamente a la obtención y al análisis de las secuencias de nucleótidos de los ge conocidos; IBM, por su parte, anuncia en un plazo de cinco años un ordenador 500 veces más potente que *D Blue*, el ordenador que en mayo de 1997 derrotó a Kasparov, y acabó así con la hegemonía humana en el aje Su nombre, *Blue Gene*, su objetivo, deducir tras un año de cálculo, la conformación tridimensional de una pe proteína (de entre las decenas de miles codificadas en nuestro genoma) a partir de su secuencia de aminoácic

Pero nos encontramos sólo en los inicios (balbuceantes) de la *era genómica*.
Tras el genoma de la especie humana seguirá, por un lado, el genoma de otras especies –y será posible enton conocer, en particular, aquellos genes responsables de la especificidad humana, de nuestra peculiar manera d en el mundo» y, por otro, el genoma de los individuos, y entonces será posible, en particular, cuantificar la aportación genética a nuestra individualidad; y podremos, por fin, abordar de forma más desapasionada la vi polémica que enfrenta herencia y ambiente para explicarnos como personas.

Ley de Moore e información genómica

En cualquier caso, el volumen de datos generados por estos proyectos será inconensurable con el volumen d que generan los proyectos genómicos hoy en día, y que ya nos parece difícilmente tratable. Pero no se trata s la información de secuencia, cada experimento con matrices de DNA genera alrededeor de unos de 60 Mega de información. Cientos, miles quizá, de dichos experimentos están siendo realizados estos días; centenares miles, millones de ellos (cuando el diagnóstico molecular se generalice) se llevarán a cabo en un futuro no m lejano.
El volumen de información que genera la investigación genómica crece y continuará creciendo a una velocid vertiginosa. De hecho, lo hace a una velocidad más elevada de lo que predice la famosa ley de Moore, de acu con la cual, la capacidad de los ordenadores se duplica cada 18 meses, una tendencia que se inició a finales d años cincuenta y que dura hasta nuestros días. Parece ser que muy pocas actividades humanas crecen a un rit superior; la investigación genómica constituye una excepción: el lapso de tiempo necesario para que se dupl volumen de secuencias de nucleótidos almacenadas en las bases de datos públicas (GenBank/EMBL/DDBJ) ejemplo, es ya inferior a un año: en marzo de 1999, estas bases de datos contenían 2300 millones de nucleóti en marzo del 2000, 6100 millones, y el ritmo de crecimiento sólo hace que acelerarse.[2] Este hecho tiene implicaciones trascendentales: la información genómica crece a una velocidad muy superior a la que crecen más, a la que pueden crecer) los recursos necesarios para analizarla. Y no se trata sólo, ni quizá principalmen recursos computacionales. Se trata, sobre todo, de recursos humanos. Dicho de otra forma, la mayor limitaci convertir la avalancha de datos genómicos en conocimiento relevante sobre los procesos de la vida no reside en día, en la capacidad insuficiente de los ordenadores, sino en la escasez de científicos y técnicos formados utilización y el desarrollo de herramientas computacionales para el análisis de esos datos genómicos. Una es que sólo ha hecho que acentuarse en los últimos cinco años y para la que no se vislumbra una solución a cor plazo. Sólo hace falta hojear las páginas de anuncios de trabajo en *Science* o *Nature* para darse cuenta de la presencia creciente de anuncios en bioinformática o biología computacional, tanto en el sector público como privado.
Ante tal situación, uno de los retos de la bioinformática es el desarrollo de métodos que permitan integrar los genómicos –de secuencia, de expresión, de estructura, de interacciones, etc.– para explicar el comportamient global de la célula viva, minimizando la intervención humana. Dicha integración, sin embargo, no puede producirse sin tener en cuenta el conocimiento acumulado durante decenas de años, producto de la investiga miles de científicos, y que se encuentra recogido en millones de comunicaciones científicas. En este sentido, ya investigando en el desarrollo de métodos para correlacionar automáticamente datos genómicos con la información recogida en artículos científicos archivados en bases de datos como Medline.[3] El objetivo, hasta punto, es producir conocimiento científico autónomo. Asistimos, pues en la bioinformática, a los primeros in de autonomizar parte de la investigación científica del ser humano. Para algunos la perspectiva puede ser inquietante, pero es posible que la investigación científica, como ha ocurrido ya con el ajedrez, deje de ser u prerrogativa humana.
Para finalizar, querría enfatizar que con la genómica la importancia de la computación en la biología no prov sólo del hecho del enorme volumen y la complejidad de los datos que las tecnologías desarrolladas alrededo esta disciplina generan y que hacen imprescindible la utilización del ordenador, como ocurre hoy en día con disciplinas científicas. Con la genómica, la relación entre biología y computación no se fundamenta sólo en «cantidad» de los datos, sino que se establece de manera más íntima y radicalmente distinta, a partir de la naturaleza de la información genómica primaria: la secuencia de nucleótidos del DNA y la secuencia de aminoácidos de las proteínas. La peculiar naturaleza de esta información (secuencias de símbolos) la hace particularmente apropiada al análisis computacional. El hecho de que las secuencias sean portadoras de una cantidad de información, en particular la que se deriva del hecho de que secuencias similares exhiben usualm

una función y una historia similares, hace este análisis excepcionalmente relevante. En este sentido, cabe des que una de las técnicas más fructíferas utilizadas en los laboratorios de biología molecular durante la década noventa es la técnica puramente computacional. Dicha técnica consiste en comparar la secuencia de un nuev con la secuencia de los genes ya conocidos depositados en las bases de datos, con el objetivo de inferir la funcionalidad del nuevo gen a partir de la funcionalidad de los genes conocidos con los que el nuevo gen exl similaridad en su secuencia. El artículo en el que se describe el programa informático subyacente a esta técn ha convertido en el artículo más citado en biología durante la década de los noventa.[4] En genómica, los ordenadores, pues, no sirven sólo para modelizar la realidad, sino también para observarla, analizarla e interpretarla. Es decir, a diferencia de la modelización matemática tradicional en biología, la realidad ha de s menudo (extraordinariamente) simplificada para construir modelos simbólicos susceptibles de ser tratados matemática y computacionalmente. En genómica, la realidad es intrínsecamente simbólica y el ordenador es instrumento mediante el cual la realidad es observada sin intermediación. Es por ello que en genómica, la computación no es sólo una herramienta para resolver determinados problemas, sino que muchos problemas pueden ni tan sólo ser planteados sino es en términos computacionales. En definitiva, con la genómica culmi proceso de reconocimiento de que la vida tiene a escala molecular un carácter esencialmente simbólico (proc iniciado, en cierto modo, por Schroedinger, quien años antes del descubrimiento de la estructura molecular d DNA por Watson y Crick, aventuró que el DNA debería ser un cristal aperiódico constituido por la sucesión número pequeño de elementos isoméricos, la secuencia precisa de los cuales, y no tanto sus características fisicoquímicas, es la responsable de su funcionalidad). Ahora sabemos que a escala molecular los procesos d vida son computaciones, en un sentido casi paradigmático, de esa secuencia. Es por ello que, en palabras de Maddox, antiguo editor de la revista *Nature*, «biología y computación, ya interdependientes, van a permanec inextricablemente unidas».

Bibliografía

(1) http://www.no.ibm.com/nyheter/des99/bluegene.html
(2) http://www.ebi.ac.uk (European Bioinformatics Institute)
(3) C. Blaschke, M.A. Andrade, C. Ouzonis and A. Valencia (1999)
Automatic extraction of biological information from scientific text: protein-protein interactions. "Proceeding the Seventh International Conference on Intelligent Systems for Molecular Biology", T. Lengauer, R. Schne Bork, D. Brutlag, J. Glasgow, H-W. Mewes and R. Zimmer. pp 60-67.
(4) S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. Lipman (1990) Basic Local Alignment Search T "Journal of Molecular Biology" 215:403--410.

Roderic Guigó Serra

Doctor en biología por la Universidad de Barcelona (UB). En el Departamento de Estadística de esta univers trabajó en el desarrollo de modelos matemáticos y computacionales en genética de poblaciones y ecología evolutiva. Desde 1989 hasta 1991 trabajó en el Molecular Biology Computer Research Resource en el Dana Cancer Institute de la Universidad de Harvard. En 1991, trabajó en el BioMolecular Engineering Research C en la Universidad de Boston. Durante estos años su investigación se centró en el campo del análisis de secue En 1992 se trasladó a Los Alamos National Laboratory; estudió básicamente los problemas relacionados coı análisis de los genomas. Desde 1994 es investigador en el Instituto Municipal de Investigación Médica (IMI Barcelona. Desde 1994 hasta 1999 fue profesor asociado en la UB. Desde 1999 es profesor asociado en la Universitat Pompeu Fabra. Cuenta con numerosos trabajos publicados en revistas de ámbito internacional.

rguigo@imim.es
http://www1.imim.es/~rguigo